Structural Deep Learning in Conditional Asset Pricing

Jianqing Fan, Zheng Tracy Ke, Yuan Liao, Andreas Neuhierl

Discussion by Andrew Patton

Duke University

ABFR webinar, March 2022

What does a neural network estimate?

- (Deep) neural networks are "universal approximators," which means they can consistently estimate unknown functions that are sufficiently smooth.
  - The authors note that DNNs are less sensitive to tuning parameters than kernel-based methods. (Jianqing is an expert in both, so this is high praise for DNNs!)

- Estimate a mean function using the *cross-section* of returns at time t:

$$\hat{m}_t(\cdot) = \arg \min_{m \in \mathcal{M}_{J,L}} \sum_{i=1}^{N} \left(r_{it} - m\left(\mathbf{x}_{i,t-1}\right)\right)^2$$

- Without any economic theory, can hope that $\mathbb{E}\left[r_{it} | \mathbf{x}_{i,t-1}\right] \approx \hat{m}_t\left(\mathbf{x}_{i,t-1}\right)$.
  - Papers in the fast-growing ML+finance literature have shown that such models produce better forecasts than standard linear models.

★ But asset pricing is done on a *panel* of asset returns. How does $\hat{m}_t\left(\mathbf{x}_{i,t-1}\right)$ relate to standard factor models for asset pricing?

A factor model for asset returns

- Consider a standard factor model for asset returns

$$r_{it} = \alpha_i + \boldsymbol{\beta}_i^{\mathsf{T}} \mathbf{f}_t + u_{it}$$

where $\mathbf{f}_t$ is a vector of K factors, with mean $\boldsymbol{\lambda}$.

A factor model for asset returns

- Consider a standard factor model for asset returns

$$r_{it} = \alpha_i + \boldsymbol{\beta}_i^{\mathsf{T}} \mathbf{f}_t + u_{it}$$

  where $\mathbf{f}_t$ is a vector of K factors, with mean $\boldsymbol{\lambda}$.

- Next consider a *conditional* version of this model

$$r_{it} = \underbrace{\alpha_{i,t-1}}_{\text{mispricing}} + \underbrace{\boldsymbol{\beta}_{i,t-1}^{\mathsf{T}} \boldsymbol{\lambda}_{t-1}}_{\text{risk premia}} + \boldsymbol{\beta}_{i,t-1}^{\mathsf{T}} \underbrace{(\mathbf{f}_t - \boldsymbol{\lambda}_{t-1})}_{\text{factor shock}} + \underbrace{u_{it}}_{\text{idio. shock}}$$

A factor model for asset returns

- Consider a standard factor model for asset returns

$$r_{it} = \alpha_i + \boldsymbol{\beta}_i^{\mathsf{T}} \mathbf{f}_t + u_{it}$$

  where $\mathbf{f}_t$ is a vector of K factors, with mean $\boldsymbol{\lambda}$.

- Next consider a *conditional* version of this model

$$r_{it} = \underbrace{\alpha_{i,t-1}}_{\text{mispricing}} + \underbrace{\boldsymbol{\beta}_{i,t-1}^{\mathsf{T}} \boldsymbol{\lambda}_{t-1}}_{\text{risk premia}} + \boldsymbol{\beta}_{i,t-1}^{\mathsf{T}} \underbrace{(\mathbf{f}_t - \boldsymbol{\lambda}_{t-1})}_{\text{factor shock}} + \underbrace{u_{it}}_{\text{idio. shock}}$$

- The first two terms contribute to the expected return, the latter two only to variation.

Interpreting the NN output using a factor model

- The conditional factor model aids the interpretation of the output of the NN:

$$\mathbb{E}\left[r_{it}|\mathbf{x}_{i,t-1},\mathbf{f}_t\right] \quad = \quad \alpha_{t-1}\left(\mathbf{x}_{i,t-1}\right)+\boldsymbol{\beta}_{t-1}^{\intercal}\left(\mathbf{x}_{i,t-1}\right)\mathbf{f}_t \quad \approx \quad \hat{m}_t\left(\mathbf{x}_{i,t-1}\right)$$

Interpreting the NN output using a factor model

- The conditional factor model aids the interpretation of the output of the NN:

$$\mathbb{E}\left[r_{it}|\mathbf{x}_{i,t-1},\mathbf{f}_t\right] \quad = \quad \alpha_{t-1}\left(\mathbf{x}_{i,t-1}\right)+\boldsymbol{\beta}_{t-1}^{\intercal}\left(\mathbf{x}_{i,t-1}\right)\mathbf{f}_t \quad \approx \quad \hat{m}_t\left(\mathbf{x}_{i,t-1}\right)$$

  - By estimating the NN period-by-period, the *factors* are also present in the estimated conditional mean.

Interpreting the NN output using a factor model

- The conditional factor model aids the interpretation of the output of the NN:

$$\mathbb{E}\left[r_{it}|\mathbf{x}_{i,t-1},\mathbf{f}_t\right] \quad = \quad \alpha_{t-1}\left(\mathbf{x}_{i,t-1}\right)+\boldsymbol{\beta}_{t-1}^{\intercal}\left(\mathbf{x}_{i,t-1}\right)\mathbf{f}_t \quad \approx \quad \hat{m}_t\left(\mathbf{x}_{i,t-1}\right)$$

  - By estimating the NN period-by-period, the *factors* are also present in the estimated conditional mean.

- The authors show how to identify all three terms that appear in the RHS.

Step 1: The conditional mean

- The first step in the estimation procedure is to use cross-sectional DNNs to flexibly estimate the conditional mean *for each period* t

$$\hat{m}_t(\cdot) = \arg \min_{m \in \mathcal{M}_{J,L}} \sum_{i=1}^{N} \left(r_{it} - m\left(\mathbf{x}_{i,t-1}\right)\right)^2$$

- The authors use a feedforward neural network
  - 3 layers
  - Number of neurons is 16, 8, 4
  - Learning rate is 0.1
  - "ReLu" activation function: $h(\mathbf{x}) = \max(\mathbf{x}, 0)$

Step 2: Estimating the betas

- Recall the assumed structure for the true mean:

$$\mathbb{E}\left[r_{it}|\mathbf{x}_{i,t-1},\mathbf{f}_t\right] \quad = \quad \alpha_{t-1}\left(\mathbf{x}_{i,t-1}\right)+\beta_{t-1}^{\intercal}\left(\mathbf{x}_{i,t-1}\right)\mathbf{f}_t \quad \approx \quad \hat{m}_t\left(\mathbf{x}_{i,t-1}\right)$$

Step 2: Estimating the betas

- Recall the assumed structure for the true mean:

$$\mathbb{E}\left[r_{it}|\mathbf{x}_{i,t-1},\mathbf{f}_t\right] \quad = \quad \alpha_{t-1}\left(\mathbf{x}_{i,t-1}\right)+\beta_{t-1}^{\intercal}\left(\mathbf{x}_{i,t-1}\right)\mathbf{f}_t \quad \approx \quad \hat{m}_t\left(\mathbf{x}_{i,t-1}\right)$$

- Consider "integrating out" the $\mathbf{f}_t$ variable:

$$\begin{aligned}\mathbb{E}\left[\ \mathbb{E}\left[r_{it}|\mathbf{x}_{i,t-1},\mathbf{f}_t\right]\ \middle|\ \mathbf{x}_{i,t-1}\right] &= \alpha_{t-1}\left(\mathbf{x}_{i,t-1}\right)+\beta_{t-1}^{\intercal}\left(\mathbf{x}_{i,t-1}\right)\mathbb{E}\left[\mathbf{f}_t|\,\mathbf{x}_{i,t-1}\right] \\ &\approx \mathbb{E}\left[\ \hat{m}_t\left(\mathbf{x}_{i,t-1}\right)\ \middle|\ \mathbf{x}_{i,t-1}\right]\end{aligned}$$

Step 2: Estimating the betas

- Recall the assumed structure for the true mean:

$$\mathbb{E}\left[r_{it}|\mathbf{x}_{i,t-1},\mathbf{f}_t\right] \quad = \quad \alpha_{t-1}\left(\mathbf{x}_{i,t-1}\right)+\beta_{t-1}^{\intercal}\left(\mathbf{x}_{i,t-1}\right)\mathbf{f}_t \quad \approx \quad \hat{m}_t\left(\mathbf{x}_{i,t-1}\right)$$

- Consider "integrating out" the $\mathbf{f}_t$ variable:

$$\begin{aligned}\mathbb{E}\left[\ \mathbb{E}\left[r_{it}|\mathbf{x}_{i,t-1},\mathbf{f}_t\right]\ \middle|\ \mathbf{x}_{i,t-1}\right] &= \alpha_{t-1}\left(\mathbf{x}_{i,t-1}\right)+\beta_{t-1}^{\intercal}\left(\mathbf{x}_{i,t-1}\right)\mathbb{E}\left[\mathbf{f}_t|\,\mathbf{x}_{i,t-1}\right] \\ &\approx \mathbb{E}\left[\ \hat{m}_t\left(\mathbf{x}_{i,t-1}\right)\ \middle|\ \mathbf{x}_{i,t-1}\right]\end{aligned}$$

- If this can be estimated, then the difference knocks out alpha:

$$\mathbb{E}\left[r_{it}|\mathbf{x}_{i,t-1},\mathbf{f}_t\right]-\mathbb{E}\left[\mathbb{E}\left[r_{it}|\mathbf{x}_{i,t-1},\mathbf{f}_t\right]|\,\mathbf{x}_{i,t-1}\right]=\beta_{t-1}^{\intercal}\left(\mathbf{x}_{i,t-1}\right)\left(\mathbf{f}_t-\mathbb{E}\left[\mathbf{f}_t|\,\mathbf{x}_{i,t-1}\right]\right)$$

Step 2: Estimating the betas, cont'd

- The authors make the reasonable assumption that the alpha and beta functions (and thus the entire mean function) is *slowly-moving* across time.

Step 2: Estimating the betas, cont'd

- The authors make the reasonable assumption that the alpha and beta functions (and thus the entire mean function) is *slowly-moving* across time.
  - Similar to Ang and Kristensen (2012, JFE).

Step 2: Estimating the betas, cont'd

- The authors make the reasonable assumption that the alpha and beta functions (and thus the entire mean function) is *slowly-moving* across time.
  - Similar to Ang and Kristensen (2012, JFE).
- Can then estimate $\mathbb{E}\left[\ \hat{m}_t\left(\mathbf{x}_{i,t-1}\right)\ \middle|\ \mathbf{x}_{i,t-1}\right]$ using kernel smoothing:

$$\bar{m}_{i,t} = \frac{1}{Th}\sum_{s=1}^{T} \hat{m}_s\left(\mathbf{x}_{i,t-1}\right) K\left(\frac{s-t}{Th}\right)$$

Step 2: Estimating the betas, cont'd

- The authors make the reasonable assumption that the alpha and beta functions (and thus the entire mean function) is *slowly-moving* across time.
  - Similar to Ang and Kristensen (2012, JFE).
- Can then estimate $\mathbb{E}\left[\ \hat{m}_t\left(\mathbf{x}_{i,t-1}\right)\ \middle|\ \mathbf{x}_{i,t-1}\right]$ using kernel smoothing:

$$\bar{m}_{i,t} = \frac{1}{Th}\sum_{s=1}^{T} \hat{m}_s\left(\mathbf{x}_{i,t-1}\right) K\left(\frac{s-t}{Th}\right)$$

- Combining with the first step the authors have

$$\hat{m}_t\left(\mathbf{x}_{i,t-1}\right) - \bar{m}_{i,t} \approx \boldsymbol{\beta}_{t-1}^{\intercal}\left(\mathbf{x}_{i,t-1}\right)\left(\mathbf{f}_t - \mathbb{E}\left[\mathbf{f}_t \middle| \mathbf{x}_{i,t-1}\right]\right)$$

Step 2: Estimating the betas, cont'd

- The authors make the reasonable assumption that the alpha and beta functions (and thus the entire mean function) is *slowly-moving* across time.
  - Similar to Ang and Kristensen (2012, JFE).
- Can then estimate $\mathbb{E}\left[\ \hat{m}_t\left(\mathbf{x}_{i,t-1}\right)\ \middle|\ \mathbf{x}_{i,t-1}\right]$ using kernel smoothing:

$$\bar{m}_{i,t} = \frac{1}{Th}\sum_{s=1}^{T} \hat{m}_s\left(\mathbf{x}_{i,t-1}\right) K\left(\frac{s-t}{Th}\right)$$

- Combining with the first step the authors have

$$\hat{m}_t\left(\mathbf{x}_{i,t-1}\right) - \bar{m}_{i,t} \approx \boldsymbol{\beta}_{t-1}^{\intercal}\left(\mathbf{x}_{i,t-1}\right)\left(\mathbf{f}_t - \mathbb{E}\left[\mathbf{f}_t \middle| \mathbf{x}_{i,t-1}\right]\right)$$

- The authors then use local PCA (Fan, et al., 2016, AoS) to estimate the $(N \times K)$ matrix of betas, $\mathbf{G}_{\beta,t-1}$, using the first K eigenvalues of

$$\frac{1}{Th}\sum_{s=1}^{T}\left(\widehat{\mathbf{m}}_s\left(\mathbf{X}_{t-1}\right) - \overline{\mathbf{m}}_t\right)\left(\widehat{\mathbf{m}}_s\left(\mathbf{X}_{t-1}\right) - \overline{\mathbf{m}}_t\right)^{\intercal} K\left(\frac{s-t}{Th}\right)$$

Step 3: Recovering the factors, alphas, risk premia

- Using the same local PCA approach, the factors can be recovered:

$$\widehat{\mathbf{f}}_t = \widehat{\mathbf{G}}_{\beta,t-1}^{\intercal} \left(\widehat{\mathbf{m}}_t \left(\mathbf{X}_{t-1} \right) - \overline{\mathbf{m}}_t \right)$$

- Then the risk premia can be recovered period-by-period

$$\widehat{\boldsymbol{\lambda}}_t = \frac{1}{N} \widehat{\mathbf{G}}_{\beta,t-1}^{\intercal} \overline{\mathbf{m}}_t$$

- And finally the alphas can be recovered:

$$\widehat{\mathbf{G}}_{\alpha,t-1} = \overline{\mathbf{m}}_t - \widehat{\mathbf{G}}_{\beta,t-1}^{\intercal} \widehat{\boldsymbol{\lambda}}_t$$

- Nice!

Econometric theory

- This approach has several intricate steps, and the authors carefully characterize the properties of these estimators.
- They show that the estimation error in each of these steps is bounded in probability, with rates that depend on:

1. φ_T, a measure of the approximation error from the DNN
   - Goes to zero as $N \longrightarrow \infty$, at a rate depending on how hard the function is to approximate
2. δ_T, a measure of the complexity of the DNN function space
   - Need the complexity to grow, but slower than $\log(NT)/N$
3. η_T, a familiar term coming from the kernel smoothing step
   - Need the bandwidth to shrink but slower than $1/T$

Main empirical results

1. In sample, around 95% of return variation is explained by risk exposures
   - Around 80% is from the factor realization, 20% from variation in risk premia
   - The remaining 5% of return variation is due to mispricing
2. Pricing errors appear to be larger for small-cap firms
3. Out-of-sample forecasts can be improved (by 50-65%) by eliminating the part of the DNN forecast attributable to trying to forecast factor realizations
   - Factor returns are almost white noise $\Rightarrow$ very hard to forecast

Summary

- I really enjoyed reading this paper and thinking about the methods proposed.
- Below are some questions that I would be curious to hear the authors' thoughts on.

Sensitivity to the number of assumed factors

- The authors find the intriguing result that about 5% of the in-sample variation in returns can be attributed to mispricing.
- But recall the pricing model:

$$r_{it} = \underbrace{\alpha_{i,t-1}}_{\text{mispricing}} + \underbrace{\beta_{i,t-1}^{\intercal}\boldsymbol{\lambda}_{t-1}}_{\text{risk premia}} + \beta_{i,t-1}^{\intercal}\underbrace{(\mathbf{f}_t - \boldsymbol{\lambda}_{t-1})}_{\text{factor shock}} + \underbrace{u_{it}}_{\text{idio. shock}}$$

- If the number of latent factors is chosen too small, the amount of apparent mispricing will be large.

★ How do the headline numbers change when the number of factors varies from 3 to 5 to, say, 7?

  - How many factors are used in the current empirical analysis?

Imposing and exploiting smoothness?

- The authors use a DNN to estimate, separately for each period, $\hat{m}_t(\cdot)$.
  - This is a pure *cross-sectional* approach.
  - To obtain the components α_{t-1} and β_{t-1} the authors use local smoothing.

$$\bar{m}_{i,t} = \frac{1}{Th} \sum_{s=1}^{T} \left(\arg \min_{m \in \mathcal{M}_{J,L}} \sum_{i=1}^{N} \left(r_{it} - m\left(\mathbf{x}_{i,t-1}\right) \right)^2 \right) K\left(\frac{s-t}{Th} \right)$$

★ Is it feasible to use a *panel* approach to combine these steps? Eg, optimize:

$$\tilde{m}_t(\cdot) = \arg \min_{m \in \mathcal{M}_{J,L}} \frac{1}{Th} \sum_{s=1}^{T} \sum_{i=1}^{N} \left(r_{it} - m\left(\mathbf{x}_{i,t-1}\right) \right)^2 K\left(\frac{s-t}{Th} \right)$$

- This exploits the assumed smoothness in m_t, and presumably reduces estimation error. But is it doable?

Imposing *extreme* smoothness

- To better understand where the gains from the proposed method accrue, I think it would be interesting to estimate a model where the functions are imposed to be *constant* over time.

$$\tilde{m}(\cdot) = \arg \min_{m \in \mathcal{M}_{J,L}} \sum_{s=1}^{T} \sum_{i=1}^{N} \left(r_{it} - m\left(\mathbf{x}_{i,t-1}\right) \right)^2$$

  - They can still be very flexible functions of the characteristics

★ Are the gains coming from the flexible functional form that the DNN provides, or from allowing for time variation, or both?

  - The time variation in firm characteristics, as well as in the composition of firms in the sample, will still lead to time-varying averages, but by imposing constancy of the function the source of variation will be easier to determine.

Why is the in-sample R^2 not 100%?

- This paper introduced me to the "double descent" phenomenon in machine learning (Belkin et al., 2019, PNAS)
  - OOS performance improves as you go from small to medium-sized models, then deteriorates as the model gets too big, hitting the worst case when the in-sample R^2 reaches 100%.
  - But then it starts to *improve* again as the model gets even more flexible, even though the in-sample R^2 plateaus (necessarily) at 100%!
  - It's a fascinating feature, also illustrated in this paper.

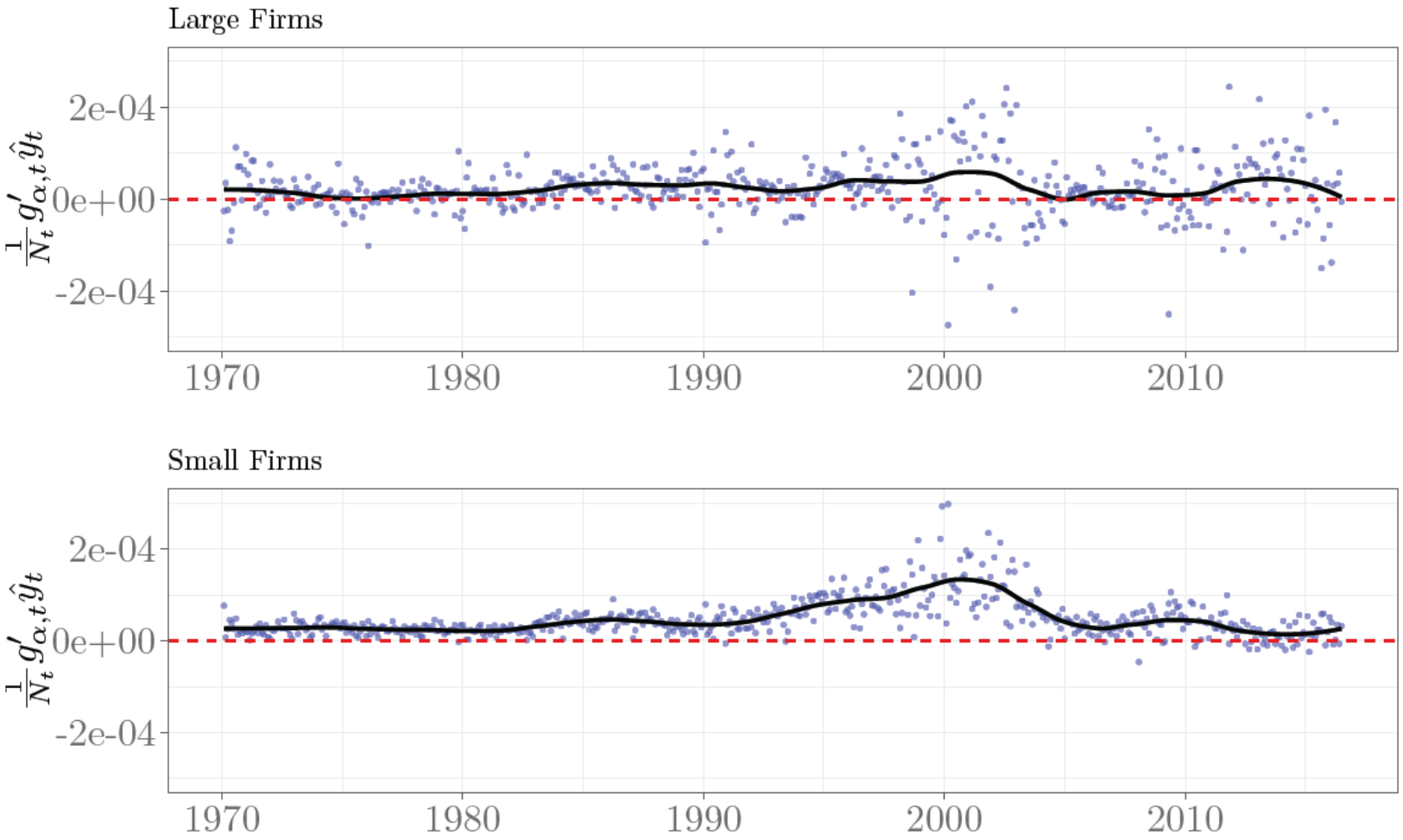
★ But in Table 1 the in-sample R^2s are a "disappointing" 12-16%. Why not 100%?

Which (types of) characteristics matter?

- The authors consider a set of 62 firm characteristics across 6 categories (past returns, investment, frictions, etc.)
- Maybe this is an old-fashioned question, in the age of machine learning + big data, but I am curious to know *which* of these characteristics matter. Or at least matter the most.
- ★ Is it feasible to do the analysis with variables only from a given category, and report how the performance varies?
  - Eg, Kaniel, Lin, Pelger and Van Nieuwerburgh (2021) do this in their DNN analysis of mutual fund performance.

Imposing some more discipline on the model

Figure 2: Evolution of average pricing errors over time



Imposing some more discipline on the model

- Figure 2 seems to show that the larger firms (top 20% by market cap) do not have pricing errors.

- ★ Would the accuracy of the model be improved if this was imposed in estimation?
  - Is that even possible, in the multi-step modeling approach here?

Correlated and/or common characteristics?

- A technical assumption is that the set of characteristics, $\mathbf{x}_{it}$, is cross-sectionally *iid*.

1. Given that firms belong to industries, sharing similar features, we might expect their characteristics to be cross-sectionally correlated.

2. Also, ruling out cross-sectional correlation means we cannot consider common variables as characteristics.
   - Eg, mispricing (alpha) is greater during periods of low liquidity.
   - Obviously, this induces extreme cross-sectional dependence.

★ How hard would it be to relax this condition?

Small comment: Fat tails

- Another key technical condition is that the tails of the innovations are thin tailed ("sub Gaussian").
- Asset returns are known to exhibit fat tails, though less at the monthly frequency than at daily or intra-daily frequencies.
- ★ How hard would it be to accommodate some excess kurtosis?
  - Have you tried simulations where the shocks are, say, $t(6)$?

Summary

- A really clever paper, using a combination of:

1. Deep neural networks for conditional mean prediction,
2. Local principal components analysis, plus
3. Standard linear factor models for asset returns to extract
   - Local betas
   - Local alphas
   - Local factors

- Rigorous econometric theory and nice empirical work.

★ I enjoyed it a lot. Even more after my 2nd and 3rd readings!